View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Zagreb Repository

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tea Ungaro

KLASTERSKA ANALIZA

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, rujan 2016.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Najveća zahvala baki, roditeljima i Dariji na razumijevanju i ljubavi, bez vas ne bih bila ovdje. Puno hvala i prof. dr. sc. Jazbec na savjetima i pomoći.

Sadržaj

Uvod

Klasterska analiza je grupiranje elemenata u klastere tako da su elementi unutar svakog klastera najsličniji mogući, dok su oni izvan što različitiji. Pravo značenje sličnosti je filozofsko pitanje, pa ćemo uzeti objektivniji pristup i gledati udaljenost između dva objekta. Udaljenost možemo mjeriti na više načina. Elementi u klasterskoj analizi mogu biti različiti, npr. ljudi, biljke, stanice, plaće, mišljenja, ali bitno je kojim su karakteristikama reprezentirani. Ideja je naći prirodno grupiranje među objektima koji se proučavaju. Objekti mogu biti opisani sa skupom karakteristika ili svojom vezom s drugim objektima. Grupiranje u klastere susrećemo u svakodnevnom životu, na primjer grupa ljudi koji objeduju za istim stolom u restorani se mogu smatrati jednim klasterom, zatim artikli slične namjene koji su u trgovini posloženi u istim odjeljcima. Ciljevi grupiranja podataka u klastere su različiti. Ponekad grupiranje u klastere pokazuje unutarnju strukturu podataka (npr. kod klasteriranja gena), ponekad je samo grupiranje cilj (npr. odvajanje proizvoda u trgovini) ili priprema za druge tehnike analize podataka. Najvažniji koraci u klasterskoj analizi su odabir mjere sličnosti i algoritma klasteriranja.

Poglavlje 1

Klasterska analiza

Klasterska analiza većinom se bavi sljedećim generalnim problemom: za dani skup podataka $\mathcal{U}$ odrediti podskupove (koji se zovu klasteri) C koji su homogeni i/ili dobro separirani u odnosu na mjerene varijable. Skup svih klastera C čini klastering $\mathbf{C}$. Ovaj problem može se formulirati kao sljedeći optimizacijski problem:
odrediti klastering $\mathbf{C}$ za koji vrijedi

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

gdje je $\mathbf{C}$ klastering za dani skup podataka $\mathcal{U}$, Φ skup svih mogućih klasteringa i $P : \Phi \rightarrow \mathbb{R}$ funkcija kriterija.
Ako je skup klastering konačan, rješenje problema uvijek postoji. Međutim, budući da je taj skup obično vrlo velik nije jednostavno naći optimalno rješenje.
Klastering $\mathbf{C} = \{C_1, C_2, ..., C_k\}$ je particija skupa $\mathcal{U}$ ako

$$\bigcup_i C_i = \mathcal{U}$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

Klastering $\mathbf{H} = \{C_1, C_2, ..., C_k\}$ je hijararhija ako za svaki par klastera C_i i C_j iz $\mathbf{H}$ vrijedi

$$C_i \cap C_j \in \{C_i, C_j, \emptyset\}$$

i potpuna hijerarhija ako za svaki podatak x vrijedi $\{x\} \in \mathbf{H}$ i $\mathcal{U} \in \mathbf{H}$.

Koraci u rješavanju problema klasteriranja su:

1. Odabrati skup podataka $\mathcal{U}$.

2. Izmjeriti varijable za dani problem. Ako su varijable numeričke, u većini slučajeva bi ih se trebalo standardizirati.

3. Izabrati prikladnu mjeru sličnosti među podacima d za zadani problem i tipove varijabli.

4. Izabrati prikladan tip klasteriranja (hijerarhijsko ili nehijerarhijsko).

5. Izabrati funkciju kriterija za procjenu odabranog tipa klasteriranja.

6. Izabrati algoritam klasteriranja.

7. Odrediti klastering(e) koji optimizira(ju) izabranu funkciju kriterija sa odabranim algoritmom.

8. Ocijeniti dobivena rješenja da se vidi imaju li neku temeljnu strukturu.

Postoji dvije vrste klasteriranja: hijerarhijsko i nehijerarhijsko. Hijerarhijski algoritmi stvaraju hijerarhijsku dekompoziciju skupa podataka služeći se nekim kriterijem. Oni su ili aglomerativni ili razdjeljujući. Aglomerativni algoritmi idu od dna prema vrhu (bottom-up) i počinju sa svakim objektom u zasebnom klasteru te spajaju grupe s obzirom na udaljenost. Algoritam može stati kada su svi objekti u jednom klasteru ili u bilo kojem trenutku korisnik želi. Razdjeljujući algoritmi idu od vrha prema dnu (top-down) i služe se obrnutom strategijom. Počinju s jednim klasterom u kojem su svi objekti i zatim ih dijeli u manje, sve dok svaki objekt nije u zasebnom klasteru ili korisnik odluči stati. Nehijerarhijski algoritmi za dani skup od n podataka konstruiraju k particija, gdje svaki klaster optimizira zadani kriterij. Poželjne karakteristike algoritama za klasteriranje su nadogradivost, mogućnost rada s različitim tipovima podataka, minimalni zahtjevi o ulaznim parametrima, dobro nošenje sa šumovima i netipičnim vrijednostima (engl. *outlierima*), neosjetljivost na red unosa podataka, upotrebljivost i mogućnost interpretacije.

1.1 Mjere sličnosti

Pretpostavimo da je $\mathcal{D}$ skup podataka o n objekata. Ako su $x, y \in \mathcal{D}$, svaki od njih ima oblik $x = (x_1, x_2, ..., x_k)$, $y = (y_1, y_2, ..., y_k)$, gdje je k dimenzija, a svaki x_i i y_i, $1 \leq i \leq n$ značajka ili atribut odgovarajućeg objekta. Atributi (varijable) mogu biti numeričke (kvantitativne) i kategorijske (kvalitativne). Numeričke varijable prirodno poprimaju vrijednosti iz skupa realnih brojeva. Tipičan primjer numeričkih varijabli su tjelesna masa ili visina osobe. Među numeričkim varijablama razlikujemo diskretne i neprekidne varijable. Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti

(npr. broj bodova na testu), dok je skup mogućih vrijednosti neprekidnih numeričkih varijabli cijeli skup realnih brojeva ili neki interval (npr. tjelesna temperatura, vodostaj rijeke). Kategorijske varijable mogu biti ordinalne ili nominalne. Karakteristika ordinalnih varijabli je da se među kategorijama može uspostaviti prirodan poredak. Tipičan je primjer takve varijable stručna sprema osobe ili školske ocjene. Kod nominalnih varijabli kategorije su neuređene, npr. mjesto rođenja. Postoje i binarne varijable koje imaju točno dvije kategorije, kao što su spol ili da/ne odgovori u anketama.
Jednom kada su karakteristike podataka određene, suočeni smo s problemom pronalaska prikladnog načina za određivanje udaljenosti između dva elementa. Odabir prikladne mjere sličnosti između dva objekta je ključan za klastersku analizu. Pri odabiru mjere sličnosti u obzir se moraju uzeti njezina matematička svojstva, vrsta podataka koje treba obraditi, ponašanje te mjere u odnosu na podatke i upotreba matrice sličnosti. Sličnost se može opisati funkcijom gdje je realan broj dodijeljen svakom paru elemenata (x, y) tj. $d : (x, y) \to \mathbb{R}$. Svojstva koja mjera sličnosti mora imati su:

- $d(x, y) \geq 0$ nenegativnost
- $d(x, x) = 0$
- $d(x, y) = d(y, x)$ simetričnost

Ako mjera sličnosti zadovoljava i sljedeća dva svojstva

- $d(x, y) = 0 \Leftrightarrow x = y$
- $\forall z : d(x, y) \leq d(x, z) + d(z, y)$ nejednakost trokuta

onda se zove metrika.
Svaka metrika je mjera sličnost, ali nije nužno svaka mjera metrika.

1.1.1 Mjere sličnosti za numeričke varijable

Numeričke varijable se mogu mjeriti na različitim skalama pa ih, prije računanja mjere sličnosti tj. udaljenosti, najčešće trebamo standardizirati. Standardizacija nije obavezna i većinom se koristi da mjerene jedinice ne bi utjecale na analizu te svakoj varijabli daje zajedničko numeričko obilježje. Najčešća standardizacija je

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

gdje je x_{ij} vrijednost varijable X_j, μ_j je aritmetička sredina, a σ_j standardna devijacija varijable X_j.
Neka su x i y opisane sa n numeričkih varijabli

$$x = (x_1, x_2, ..., x_n)$$

$$y = (y_1, y_2, ..., y_n)$$

Sličnost između x i y se može računati pomoću sljedećih udaljenosti:

- Minkowski udaljenost definirana sa

$$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{\frac{1}{r}}, r > 0$$

  Svojstvo Minkowski udaljenosti: što je veći r, to je jači utjecaj velikih razlika $|x_i - y_i|$ na dobivenu udaljenost između podataka.

- Euklidska udaljenost definirana sa

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  Primijetimo da je euklidska udaljenost jednaka Minkowski udaljenosti za $r = 2$.

- Manhattan udaljenost definirana sa

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

  Primijetimo da je Manhattan udaljenost jednaka Minkowski udaljenosti za $r = 1$.

- Čebiševljeva udaljenost definirana sa

$$d(x, y) = \max_{i=1,...,n} |x_i - y_i|$$

  Primijetimo da je Čebiševljeva udaljenost jednaka Minkowski udaljenosti za $r \to \infty$.

Također je moguće koristiti Pearsonov koeficijent korelacije kao mjeru sličnosti:

$$r(x, y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \sum_{i=1}^{n} (y_i - \mu_y)^2}}$$

gdje je

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

i

$$\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Postoje mnoge druge mjere udaljenosti na $\mathbb{R}^n$ kao na primjer Mahalanobisova generalizirana udaljenost definirana sa

$$d(x, y) = (x - y)'\Sigma^{-1}(x - y)$$

gdje je Σ kovarijacijska matrica varijabli unutar klastera. Ova udaljenost uzima u obzir odnos među varijablama. Ako je Pearsonova korelacija među varijablama 0 i varijable su standardizirane, onda je Mahalanobisova udaljenost jednaka kvadratu euklidske udaljenosti.
Ako podaci sadrže samo pozitivne vrijednosti varijable, onda imamo Lance-Williamsovu udaljenost

$$d(x, y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)}$$

i Canberra udaljenost definiranu sa

$$\dot{(}x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|}$$

Obje udaljenosti su vrlo osjetljive na male vrijednosti oko 0.

1.1.2 Mjere sličnosti za kategorijske varijable

Kategorijske varijable su nominalne ili ordinalne. Karakteristika nominalnih varijabli je da su neuređene i među njima se ne može uspostaviti prirodni poredak, tj. ne može ih se staviti na neku skalu. Sličnost između objekata x i y je dana sa

$$d(x, y) = \frac{m - p}{p}$$

gdje je m broj atributa s tom vrijednosti, a p ukupan broj atributa.
Sličnost između ordinalnih varijabli može se računati slično kao za numeričke. Neka je i ordinalni atribut sa M_i stanja koja može poprimiti. Provode se sljedeći koraci:

1. Stanja M_i su uređena $[1, ..., M_i]$, pa se svaka vrijednost može zamijeniti s odgovarajućim rangom $r_i \in \{1, ..., M_i\}$

2. Svaki ordinalni atribut bi mogao imati različitu veličinu domene, pa je potrebno svaki oblik prebaciti u interval vrijednosti [0.0, 1.0]. To se dobiva sljedećom transformacijom:
$$z_i^{(j)} = \frac{r_i^{(j)} - 1}{M_i - 1}$$

3. Sličnost se može izračunati s jednom od metrika za numeričke varijable koristeći $z_i^{(j)}$

1.1.3 Mjere sličnosti za binarne (dihotomne) varijable

Mnoge mjere sličnosti su definirane za podatke opisane binarnim (dihotomnim) varijablama. Pri prikazivanju binarnih varijabli često se koristi frekvencijska tablica. Neka su x i y podaci gdje su vrijednosti n varijabli + i – dani u tablici:

	+	–
+	a	b
–	c	d

Zbroj sve četiri frekvencije jednak je broju varijabli, tj. $a+b+c+d = n$. Frekvencija a broji koliko ima pozitivnih odgovora za x i y, d broji koliko ima negativih, dok b i c broje koliko je varijabli za koje x i y imaju različite odgovore (x pozitivne i y negativne, te obratno). Neke mjere sličnosti za binarne podatke su:

- Sokal-Michener udaljenost (1958.)
$$\frac{a + d}{a + b + c + d}$$

- Prva Sokal-Sneath udaljenost (1963.)
$$\frac{2(a + d)}{2(a + d) + b + c}$$

- Rogers-Tanimoto udaljenost (1960.)
$$\frac{a + d}{a + d + 2(b + c)}$$

- Russel-Rao udaljenost (1940.)

$$\frac{a}{a+b+c+d}$$

- Jaccard udaljenost (1908.)

$$\frac{a}{a+b+c}$$

- Czekanowski udaljenost (1913.)

$$\frac{2a}{2a+b+c}$$

- Druga Sokal-Sneath udaljenost (1963.)

$$\frac{a}{a+2(b+c)}$$

- Kulczynski udaljenost (1927.)

$$\frac{a}{b+c}$$

Sve udaljenosti osim posljednje su definirane na intervalu između 0 i 1. Prve tri udaljenosti su ekvivalente. Također, Jaccard, Czekanowski i druga Sokal-Sneath udaljenost su ekvivalentne. Pojam ekvivalentnosti mjera sličnosti je bitan u klasterskoj analizi. Neke metode klasteriranja daju iste rezultate kada se koriste različite, ali ekvivalentne udaljenosti među podacima.

1.1.4 Mjere sličnosti koje prihvaćaju sve vrste podataka

Postoje mjere sličnosti koje mogu raditi sa svim vrstama podataka, kao npr.

- Gower udaljenost

$$d(x,y) = \frac{\sum_{j=1}^{n} d_{x,y}^{j}}{n}$$

gdje je za norminalne i binarne (dihotomne) varijable

$$d_{x,y}^{j} = \begin{cases} 1, & x_j = y_j \\ 0, & x_j \neq y_j \end{cases}$$

te za numeričke i ordinalne varijable

$$d_{x,y}^{j} = 1 - |x_j - y_j|$$

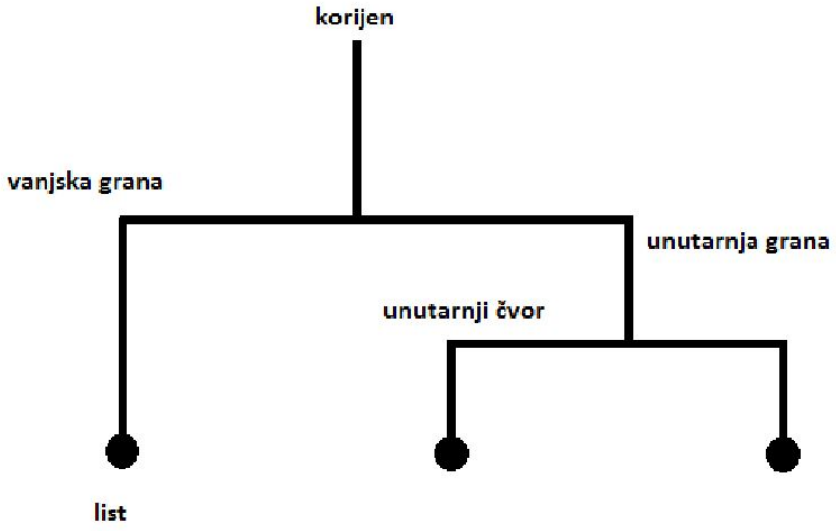
- DGower udaljesnost
$$d_2(x, y) = 1 - d(x, y)$$
gdje je $d(x, y)$ Gower udaljenost.

1.2 Algoritmi klasteriranja

Općenito, ne postoji objektivno ispravan algoritam klasteriranja. Određeni algoritam može dati dobre rezultate na jednom skupu podataka, a loše ili nikakve na drugom, ovisno o dimenziji, strukturi i vrsti podataka. Karakteristike koje bi trebao imati dobar algoritam klasteriranja su sposobnost izvođenja na velikim skupovima podataka i različitim tipovima varijabli, te minimalni zahtjevi o ulaznim parametrima. Algoritmi klasteriranja bi se trebali dobro nositi s devijacijama. Devijacije se definiraju kao objekti koji odstupaju od generalnog ponašanja podataka i odnose se kao outlieri. Isti skup podataka zadan u različitom redoslijedu u određenim algoritmima može dati drastično različite rezultate, zato je bitno da algoritmi budu neosjetljivi na redoslijed unosa podataka. Potrebno je da algoritmi klasteriranja daju korisne rezultate koje je lako interpretirati, te usporediti s unaprijed stvorenim idejama i očekivanjima.

1.2.1 Hijerarhijski algoritmi

Hijerarhijski algoritmi su bazirani na ideji da su objekti više povezani s objektima koji su im bliži nego s onima koji su im udaljeniji. Oni formiraju klastere pomoću mjere sličnosti d između svakog novog elementa i svih ostalih već prije određenih klastera. Stablo kojim se prikazuje raspored klastera nastalih hijerarhijskim klasteriranjem naziva se dendrogram. Sličnost između dva objekta u dendrogramu je reprezentirana visinom odnosno duljinom najnižeg unutarnjeg čvora kojeg dijele. Ako je dendrogram u koordinatnom sustavu, na osi ordinata vidi se visina na kojoj na pojedini klasteri spajaju, dok su na osi apscisa raspoređeni objekti. U dendrogramu se netipične vrijednosti (engl. *outlieri*) tj. podaci jako različiti od ostalih uočavaju kao izolirana grana.



Slika 1.1: Dendogram

Aglomerativni hijerarhijski algoritmi gledaju svaki objekt posebno te između svaka dva traže najbolji par koji će spojiti u klaster. To ponavljaju dok svi klasteri nisu spojeni. Pretpostavimo da su sve relevantne informacije o odnosu između n objekata iz skupa $\mathcal{U}$ u simetričnoj matrici $D = [d_{ij}]$.

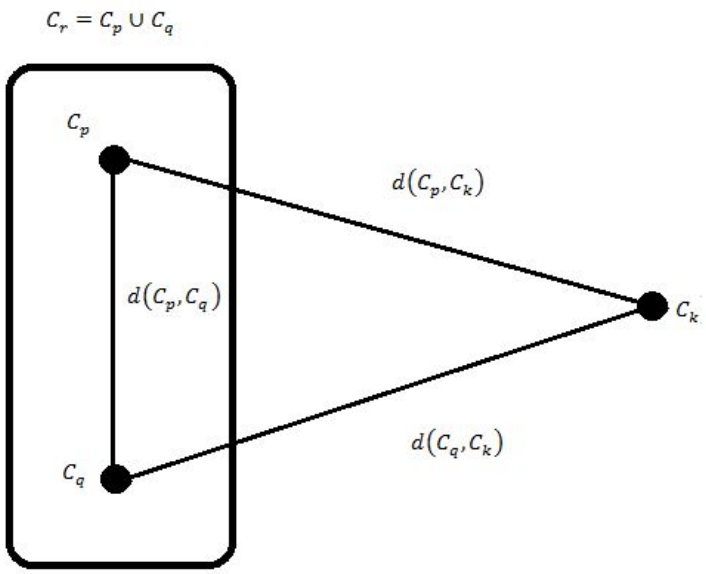
Shematski prikaz aglomeracijskog algoritma:

Svaki podatak je klaster $C_i = \{x\}, x_i \in \mathcal{U}, i = 1, 2, ..., n$;
ponavljaj dok god postoje barem dva klastera:
odredi najbliži par klastera C_p i C_q:

$$d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$$

spoji klastere C_p i C_q i stvori novi klaster $C_r = C_p \cup C_q$
zamijeni C_p i C_q sa klasterom C_r;
odredi sličnost između C_r i ostalih klastera.

Prema zadnjem koraku algoritma, trebamo odrediti sličnost d između novog klastera C_r i svih ostalih prije formiranih. To se može učiniti na mnogo načina, od kojih svaki određuje drukčiju metodu hijerarhijskog klasteriranja. Pretpostavimo da imamo tri klastera C_i,C_j i C_k u određenom koraku iteracije gornjeg algoritma. Neka su klasteri C_i i C_j najbliži. Oni



Slika 1.2: Tri klastera

su spojeni i formiraju novi klaster $C_i \cup C_j$. Metode za računanje sličnosti između novog klastera i postojećeg C_k su sljedeće:

- Metoda minimuma (engl. *Single linkage*)

  $$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

  Udaljenost između dva klastera najkraća je udaljenost između bilo kojeg elementa u prvom klasteru i bilo kojeg elementa u drugom klasteru. To je zapravo udaljenost između najsličnijih elemenata. Ova metoda najčešće stavlja elemente u dugačke i uske klastere u obliku lanaca.

- Metoda maksimuma (engl. *Complete linkage*)

  $$d(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

  Udaljenost između dva klastera najveća je udaljenost između bilo kojeg elementa u prvom klasteru i bilo kojeg elementa u drugom klasteru. To je zapravo udaljenost između najrazličitijih elemenata. Ova metoda najčešće stavlja elemente u klastere u obliku krugova.

- Metoda prosjeka (engl. *Group average*)

$$d(C_i \cup C_j, C_k) = \frac{1}{(n_i + n_j)n_k} \sum_{u \in C_i \cup C_j} \sum_{v \in C_i \cup C_j} d(u, v)$$

  pri čemu je n_i broj elemenata u klasteru C_i. Udaljenost između dva klastera prosjek je svih mogućih udaljenosti između bilo kojeg elementa u prvom i u drugom klasteru. To je najčešće korištena metoda.

- Gower metoda

$$d(C_i \cup C_j, C_k) = d^2(t_{ij}, t_k)$$

  pri čemu t_{ij} označava centar klastera $C_i \cup C_j$, a t_k centar C_k. Udaljenost između dva klastera jednaka je kvadratu udaljenosti između centara klastera.

- Ward metoda (minimum varijance)

$$d(C_i \cup C_j, C_k) = \frac{(n_i + n_j)n_k}{n_i + n_j + n_k} d^2(t_{ij}, t_k)$$

  Povezuje dva elementa u klaster tako da varijance unutar klastera bude minimalna. Kod ove metode klasteri su često u obliku elipsi.

Aglomeracijski hijerarhijski algoritmi su jednostavni i njihova rješenja se intuitivno mogu iščitati iz dendrograma, no interpretacija rezultata je subjektivna. Prednost ovog algoritma je što nije potrebno unaprijed znati broj klastera. Nedostatak je što algoritam ne može poništiti ono što je prethodno napravio.

1.2.2 Nehijerarhijski algoritmi

Kod nehijararhijskog klasteriranja svaki objekt se smješta u točno jedan od k disjunktnih klastera. Broj klastera mora biti unaprijed određen. Najpoznatiji nehijerarhijski algoritam je algoritam k-srednjih vrijednosti tj. *k-means*.

- Algoritam k-srednjih vrijednosti (engl. *k-means*) sastoji se od niza koraka:

  1. Izaberi broj klastera k.
  2. Inicijaliziraj k centara klastera (slučajnim odabirom).
  3. Svaki od n objekata pridruži najbližem centru klastera.
  4. Promijeni centre klastera pretpostavljajući da su objekti stavljeni u točne klastere.

5. Ponavljaj korake 3. i 4. sve dok niti jedan od n objekata ne promijeni svoj klaster.

K-means optimizira prosječnu udaljenost članova u istom klasteru

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|x_{ki} - x_{kj}\|^2$$

Cilj je smanjiti pogrešku SSE koja je suma kvadrata udaljenosti svake točke od pripadnog centra klastera.

$$SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_{ki} - \mu_k\|^2$$

Jedan od načina da se smanji pogreška SSE je povećanje broja klastera k. No, dobar klastering s malim k ima manju pogrešku SSE nego loš klastering s velikim brojem klastera. Centri klastera su obično srednja vrijednost objekata unutar klastera. Prednost ovog algoritma je što je jednostavan za implementaciju, te intuitivan jer optimizira sličnost unutar klastera. Nedostaci su što broj klastera mora biti unaprijed određen, loše se nosi s netipičnim vrijednostima (engl. *outlierima*), nije ga moguće primijeniti na kategorijske podatke, često završava u lokalnom optimumu te je zbog toga inicijalizacija važna.

- Algoritam k-modova (engl. *k-modes*)
  Ideja je ista kao kod algoritma k-srednjih vrijednosti, te se struktura algoritma ne mijenja. Jedina je razlika u mjeri sličnosti. Za dvije kategorijske varijable x i y mjera sličnosti se računa kao

$$d(x, y) = \sum_{i=1}^{n} \delta(x_i, y_i)$$

  gdje je

$$\delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_j \neq y_j \end{cases}$$

  Intuitivno, gornji izraz broji koliko različitih vrijednosti dva objekta imaju u pripadajućim atributima. Nemaju svi atributi istu težinu. Stoga, ako uzmemo u obzir frekvencije vrijednost u skupu podataka, mjera sličnosti izgleda ovako

$$d(x, y) = \sum_{i=1}^{n} \frac{n_{x_i} + n_{y_i}}{n_{x_i} n_{y_i}} \delta(x_i, y_i)$$

gdje su n_{x_i} i n_{y_i} broj objekata u skupu podataka s vrijednostima x_i i y_i za atribut i. Mod skupa je vrijednost koja se najčešće pojavljuje u skupu. Za skup podataka dimenzije n, svaki klaster C, $1 \leq C \leq k$, ima mod definiran s vektorom $Q^C = (x_1^C, x_2^C, \ldots, x_n^C)$. Rezultat ove metode je skup vektora Q^C koji minimiziraju izraz

$$E = \sum_{C=1}^{k} \sum_{x \in C} d(x, Q^C).$$

Poglavlje 2

Primjer

2.1 Opis podataka

Podaci na kojima ću provesti klastersku analizu su baza dijabetičara i sastoje se od 50% slučajnog uzorka stvarnih podataka dobivenih iz Sveučilišne klinike Vuk Vrhovec. Pacijenti su podijeljeni u 4 skupine: pacijenti bez dijabetesa tj. kontrolna skupina (grupa 0), pacijenti s dijebetesom tipa I (grupa 1), pacijenti s dijebetesom tipa II (grupa 2) i pacijenti sa predijabetesom (grupa 3). Varijable koje se nalaze u bazi su spol (1 ili 2), dob, HbA1c (glikolizirani hemoglobin) koji pokazuje regulaciju glikemije za 3 mjeseca, fBG (engl. *fasting blood glucosae*) glukoza natašte, ppBG (engl. *posprandial blood glucosae*) glukoza 2 sata nakon obroka, fC peptid (engl. *fasting C-peptide*) C-peptid natašte, ppC peptid (engl. *postprandial C-peptide*) C-peptid 2 sata nakon obroka, CRP (engl. *C-reactiv proteine*) upalni marker, HCY (engl. *homocysteine*) upalni marker, HDL (engl. *high-dencity lipoprotein*) jedna lipidna frakcija, TG (engl. *triglycerides*) također lipidna frakcija, Lp(a) (lipoprotein), UA (mokraćna kiselina) i indeks tjelesne mase BMI.

2.2 Analiza podataka

Za sve varijable napravljena je deskriptivna statistika. Za sve statističke analize korišten je statistički paket SAS, a za grafičke prikaze Microsoft Excel i statistički paket SAS.

	Dob					
	N	Min	Median	Max	Mean	Std
Groups						
0	24	25.00	52.50	77.00	52.58	12.34
1	26	20.00	50.00	69.00	48.04	11.87
2	189	32.00	62.00	84.00	61.48	11.00
3	21	19.00	61.00	89.00	60.57	14.34
All	260	19.00	60.00	89.00	59.24	12.30

	HbA1c					
	N	Min	Median	Max	Mean	Std
Groups						
0	23	4.70	5.30	6.00	5.42	0.34
1	26	5.30	6.50	10.30	6.92	1.44
2	187	5.10	7.10	13.20	7.36	1.43
3	21	5.20	5.70	6.50	5.78	0.36
All	257	4.70	6.80	13.20	7.01	1.47

	fBG					
	N	Min	Median	Max	Mean	Std
Groups						
0	23	4.50	5.20	6.40	5.27	0.49
1	26	4.00	7.65	16.60	8.16	3.08
2	188	4.60	8.65	20.40	9.30	2.95
3	21	4.70	5.80	8.60	6.00	0.92
All	258	4.00	7.90	20.40	8.56	3.05

Slika 2.1: Deskriptivna statistika za varijable Dob, HbA1c, fBG (ispis iz SAS-a)

	ppBG					
	N	**Min**	**Median**	**Max**	**Mean**	**Std**
Groups						
0	7	5.40	6.20	7.10	6.30	0.57
1	22	3.90	8.55	19.30	10.14	4.36
2	157	4.90	12.10	22.90	12.29	3.93
3	20	5.60	7.00	12.10	7.56	1.82
All	206	3.90	10.90	22.90	11.40	4.14

	fC_peptid					
	N	**Min**	**Median**	**Max**	**Mean**	**Std**
Groups						
0	11	0.33	0.65	0.92	0.59	0.18
1	22	0.01	0.03	1.42	0.17	0.31
2	163	0.02	0.64	2.91	0.71	0.42
3	16	0.36	0.71	1.77	0.75	0.37
All	212	0.01	0.60	2.91	0.65	0.43

	ppC_peptid					
	N	**Min**	**Median**	**Max**	**Mean**	**Std**
Groups						
0	4	0.71	2.27	2.69	1.98	0.94
1	9	0.02	0.24	3.77	0.74	1.18
2	122	0.02	1.51	4.04	1.61	0.83
3	16	0.38	2.43	6.05	2.72	1.54
All	151	0.02	1.53	6.05	1.69	1.03

Slika 2.2: Deskriptivna statistika za varijable ppBG, fC_peptid, ppC_peptid (ispis iz SAS-a)

	CRP					
	N	**Min**	**Median**	**Max**	**Mean**	**Std**
Groups						
0	23	0.30	1.40	11.50	2.09	2.31
1	26	0.20	0.85	15.30	1.62	2.94
2	178	0.10	2.45	36.10	3.75	4.79
3	19	0.30	2.90	9.20	3.31	2.73
All	246	0.10	2.00	36.10	3.33	4.36

	HCY					
	N	**Min**	**Median**	**Max**	**Mean**	**Std**
Groups						
0	21	7.70	12.00	23.40	13.52	4.19
1	26	7.70	10.90	20.90	12.02	3.63
2	183	7.90	15.90	80.10	17.07	7.27
3	20	10.30	14.75	36.90	16.60	6.92
All	250	7.70	15.00	80.10	16.21	6.93

	HDL					
	N	**Min**	**Median**	**Max**	**Mean**	**Std**
Groups						
0	24	1.04	1.48	2.27	1.46	0.35
1	26	0.99	1.66	2.08	1.59	0.32
2	189	0.43	1.25	2.43	1.31	0.36
3	21	0.80	1.43	2.46	1.49	0.40
All	260	0.43	1.34	2.46	1.37	0.37

Slika 2.3: Deskriptivna statistika za varijable CRP, HCY, HDL (ispis iz SAS-a)

	TG					
	N	Min	Median	Max	Mean	Std
Groups						
0	24	0.72	1.70	5.65	1.77	0.99
1	26	0.42	0.87	4.53	1.15	0.83
2	189	0.48	1.61	25.48	2.06	2.07
3	21	0.75	1.27	72.00	4.99	15.38
All	260	0.42	1.53	72.00	2.18	4.72

	Lp_a					
	N	Min	Median	Max	Mean	Std
Groups						
0	23	3.00	6.80	92.30	16.64	21.26
1	26	0.30	9.90	101.80	25.24	30.84
2	185	0.03	14.10	332.00	35.23	49.26
3	21	0.30	16.70	120.00	23.71	31.98
All	255	0.03	11.60	332.00	31.59	44.83

	UA					
	N	Min	Median	Max	Mean	Std
Groups						
0	23	220.00	288.00	481.00	293.91	55.13
1	26	125.00	273.00	461.00	261.04	87.43
2	188	167.00	330.00	620.00	337.57	84.91
3	20	186.00	346.50	516.00	345.05	96.36
All	257	125.00	319.00	620.00	326.50	87.24

Slika 2.4: Deskriptivna statistika za varijable TG, Lp_a, UA (ispis iz SAS-a)

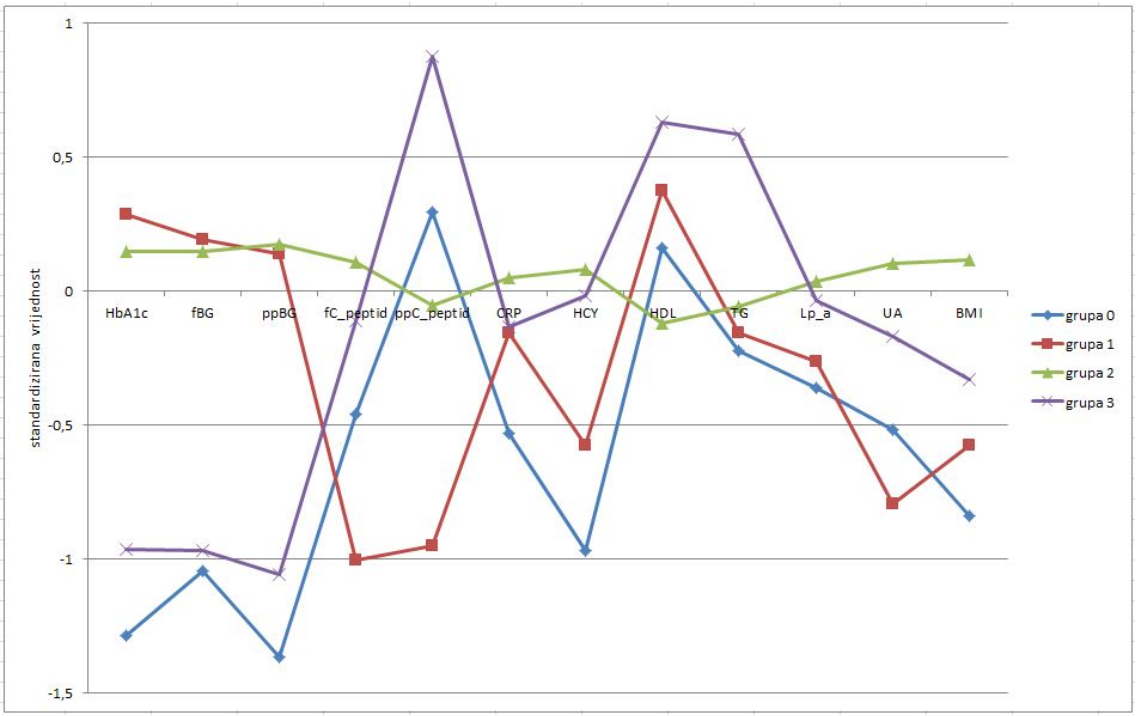
	BMI					
	N	Min	Median	Max	Mean	Std
Groups						
0	24	21.08	25.87	45.90	28.63	6.36
1	26	15.42	23.97	39.47	24.39	5.19
2	189	21.60	29.64	66.06	30.77	6.44
3	21	22.49	29.22	38.58	29.28	4.61
All	260	15.42	28.84	66.06	29.81	6.46

Slika 2.5: Deskriptivna statistika za varijablu BMI (ispis iz SAS-a)

Podaci su najprije standardizirani. Procedura `stdize` standardizira podatke iz baze `data` i sprema ih u bazu `out`. Naredba `nomiss` omogućava da se pri standardizaciji podataka preskoče vrijednosti koje nedostaju.

SAS kod:

```
proc stdize data=tea out=standa method=std nomiss;
var HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a UA BMI;
run;
```



Slika 2.6: Aritmetičke sredine stvarnih grupa (0-3)

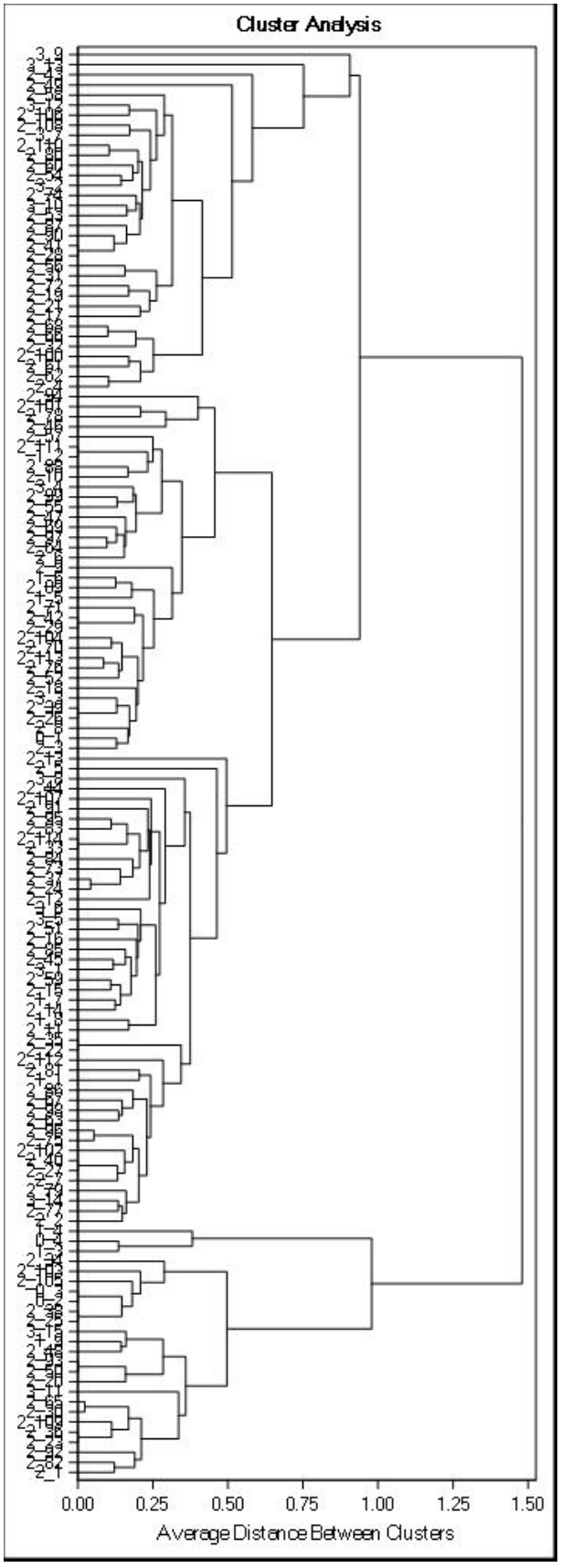
Na slici 2.6 prikazane su aritmetičke sredine mjerenih varijabli po stvarnim grupama. Mjerene varijable za grupe 0 i 3 ponašaju se približno slično, što ne čudi uzevši u obzir da su to kontrolna grupa i međugrupa tj. pacijenti s predijabetesom. Grupe 1 i 2 imaju vrlo slične vrijednosti za hemoglobin (HbA1c), glukozu natašte (fBG) i 2 sata nakon obroka (ppBG) te trigliceride, a veća razlika među aritmetičkim sredinama mjerenih varijabli vidi se kod fC i ppC peptida. Razlika između prvog i drugog tipa dijabetesa mogla bi se vidjeti i kod CRP-a, homocisteina (HCY), lipoproteina, mokraćne kiselina (UA) i indeksa tjelesne mase koji su kod tipa II iznadprosječni, a kod tipa I ispodprosječni.

Nakon standardizacije podataka potrebno je odlučiti koju mjeru sličnosti upotrijebiti, te pomoću nje napraviti matricu sličnosti koja se zatim unosi u algoritam klasteriranja. Prvo sam koristila hijerarhijske algoritme uz različite mjere udaljenosti. Kada je izabrana euklidska mjera sličnosti i metoda prosjeka, SAS kod je sljedeći:

```
proc distance data=standa out=Dist nostd method=euclid;
var interval(Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a
UA BMI);
id Groups;
```

```
run;
ods graphics on;
proc cluster data=Dist (type=distance) outtree=tree method=average
plots=dendrogram;
id Groups;
run;
```

Procedura `distance` radi matricu sličnosti podataka iz baze `data` (u ovom slučaju to su prije standardizirani podaci), te je sprema u bazu `out`. Naredba `nostd` onemogućava standardizaciju, jer je to napravljeno u koraku ranije, dok se naredbom `method` odabire mjera sličnosti. Procedura `cluster` radi klastersku analizu podataka iz skupa `data` za koje je naglašeno kojeg su tipa, tj. da su u matrici sličnosti. Naredbom `method` odabire se metoda hijerarhijskog klasteriranja. Rezultat klasterske analize bit će u obliku dendograma.

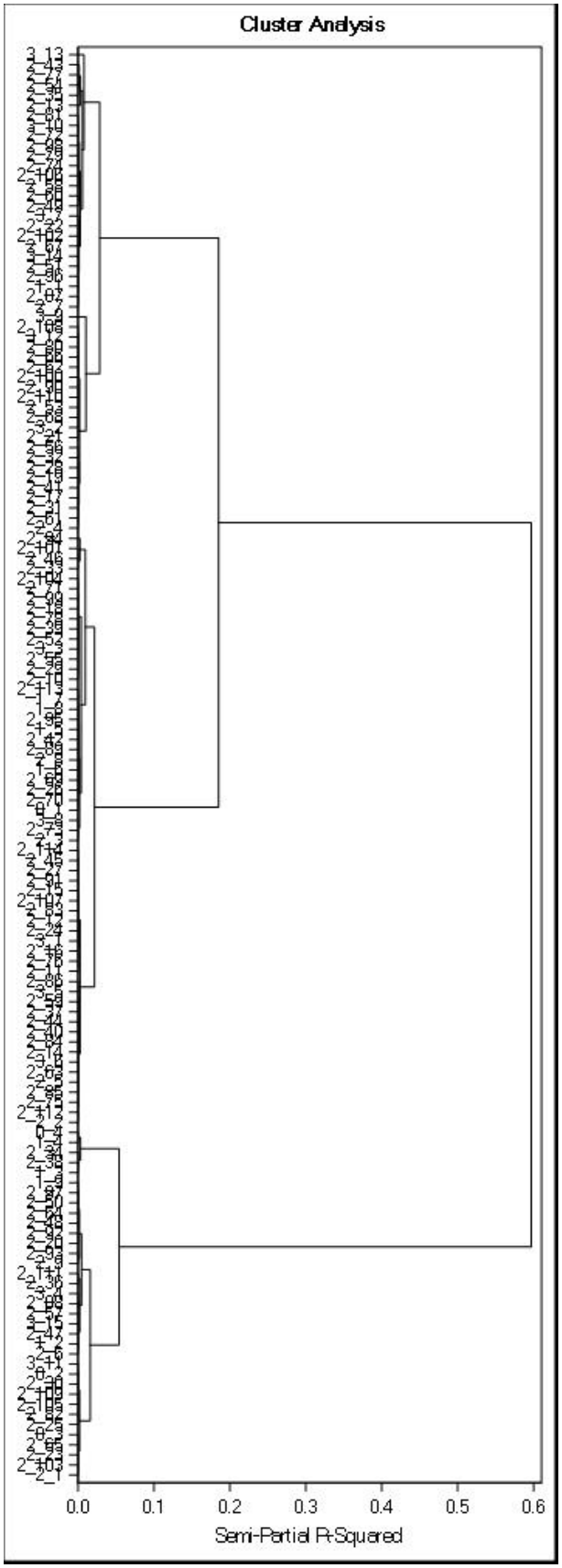


Slika 2.7: Dendogram klasterske analizu za euklidsku mjeru udaljenosti i metodu prosjeka

Dendogram klasterske analize za euklidsku mjeru udaljenosti i metodu prosjeka prikazan je na slici 2.7. Vidi se da se podaci grupiraju u četiri klastera, međutim ti klasteri ne odgovaraju stvarnoj podjeli pacijenata po grupama.

Kada je izabrana Gowerova mjera sličnosti i Wardova metoda klasteriranja, SAS kod je sljedeći:

```
proc distance data=standa out=Dist nostd method=gower;
var interval(Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a
UA BMI);
id Groups;
run;
ods graphics on;
proc cluster data=Dist (type=distance) outtree=tree method=ward
plots=dendrogram;
id Groups;
run;
```

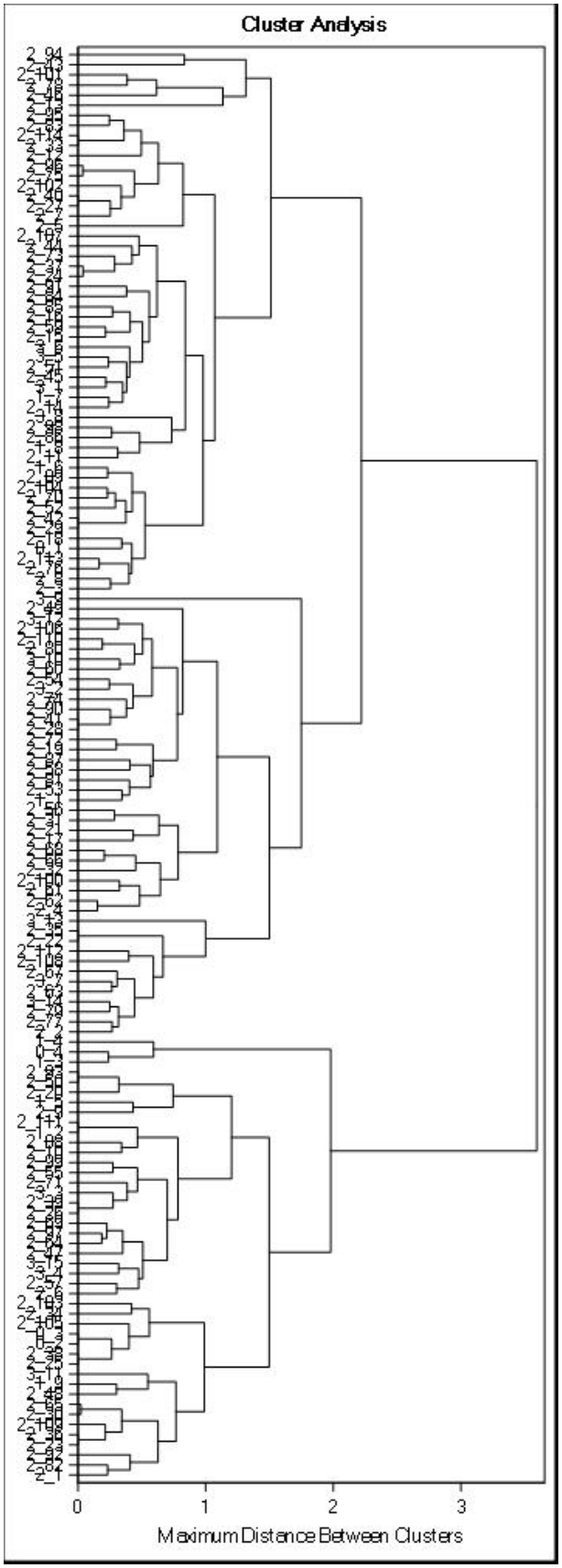


Slika 2.8: Dendogram klasterske analize za Gowerovu mjeru udaljenosti i Wardovu metodu

Dendogram klasterske analize za Gowerovu mjeru udaljenosti i Wardovu metodu klasteriranja prikazan je na slici 2.8. Vidi se da se podaci grupiraju u tri klastera, međutim ti klasteri ne odgovaraju stvarnoj podjeli pacijenata po grupama.

Kada je izabrana Manhattan mjera sličnosti (u SAS-u ima naziv cityblock) i metoda maksimuma, SAS kod je sljedeći:

```
proc distance data=standa out=Dist nostd method=cityblock;
var interval(Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a
UA BMI);
id Groups;
run;
ods graphics on;
proc cluster data=Dist (type=distance) outtree=tree method=complete
plots=dendrogram;
id Groups;
run;
```



Slika 2.9: Dendogram klasterske analize za Manhattan mjeru udaljenosti i metodu maksimuma

Dendogram klasterske analize za Manhattan mjeru udaljenosti i metodu maksimuma prikazan je na slici 2.9. Vidi se da su podaci također grupirani u tri klastera, međutim ti klasteri ne odgovaraju podjeli pacijenata po grupama.

Nakon hijerarhijskih algoritama koristila sam nehijerarhijski algoritam k-srednjih vrijednosti, pri čemu sam odabrala da je broj klastera 4, jer su u toliko grupa podijeljeni pacijenti. Pripadni SAS kod je:

```
proc fastclus data=standa out=Clust maxclusters=4 nomiss maxiter=300;
var Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a UA BMI;
run;
```

Procedura `proc fastclus` provodi *k-means* algoritam sa podacima iz baze `data` koji su prije standardizirani i sprema ih u bazu `out`. Naredbom `maxclusters` unaprijed se odabire broj klastera, dok se s `nomiss` preskaču podaci koji nedostaju.

Cluster Means

Cluster	Dob	HbA1c	fBG	ppBG	fC_peptid	ppC_peptid	CRP	HCY
1	36.52631579	-0.37420290	-0.48739961	-0.34057346	0.02403838	0.08730065	-0.42135597	-0.49305787
2	63.65384615	0.04622259	-0.05039683	-0.00115154	0.24024441	0.30089201	0.00958893	0.05774027
3	51.92500000	0.26146802	0.21822595	0.04613276	-0.25770673	-0.33534852	0.24283981	-0.28269846
4	74.58064516	-0.18556260	0.10168355	0.15114405	-0.08519902	-0.12552116	-0.07117592	0.57011433

Cluster Means

Cluster	HDL	TG	Lp_a	UA	BMI
1	-0.19256841	-0.10444144	-0.14716185	-0.36829039	-0.43476751
2	0.15876486	0.18262262	-0.00932892	0.11173627	0.10060834
3	-0.16034342	-0.07736806	-0.05121104	-0.17307652	0.13935955
4	0.05860530	-0.14249248	0.17192325	0.26162231	-0.08211074

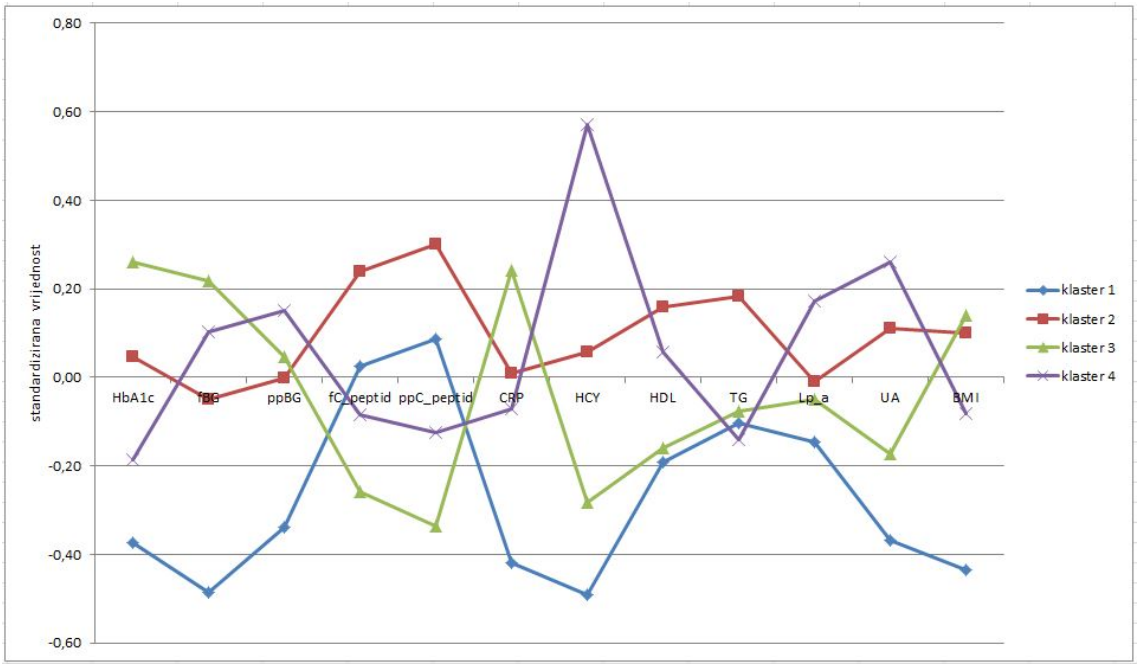
Cluster Standard Deviations

Cluster	Dob	HbA1c	fBG	ppBG	fC_peptid	ppC_peptid	CRP	HCY
1	6.449534187	0.901690039	0.608664695	0.953284216	0.936411228	1.264521804	0.311272316	0.572406213
2	3.247228776	1.125069538	1.009592497	0.987411560	1.056394189	0.947904691	1.283127622	0.750323639
3	3.452145375	0.951002896	1.053698371	1.039157598	1.012157747	0.891235354	0.973752237	0.550933646
4	4.014757722	0.815777667	1.035637072	0.996780464	0.865142041	0.923324979	0.643304915	1.596740548

Cluster Standard Deviations

Cluster	HDL	TG	Lp_a	UA	BMI
1	0.718874347	0.204674953	0.789163760	0.758964592	1.042459697
2	1.176758069	1.628330149	1.189830907	1.104354199	1.154418746
3	0.952384026	0.210202802	0.820510556	0.826862176	0.932053138
4	0.867000010	0.127533690	0.998264130	1.083532333	0.694942603

Slika 2.10: Aritmetičke sredine i standardne devijacije varijabli po klasterima (ispis iz SAS-a)



Slika 2.11: Aritmetičke sredine 4 dobivena klastera po analiziranim varijablama

Na slici 2.11 vidi se da su u klasteru 1 hemoglobin (HbA1c), glukoza natašte (fBG) i 2 sata nakon obroka (ppBG) znatno niži nego u ostala tri dobivena klastera, dok klaster 3 ima niži fC i ppC peptid od ostalih. Varijabla HCY radi veliku razliku između sva četiri klastera. Trigliceridi (TG) su kod klastera 1, 3 i 4 vrlo slični, dok je mokraćna kiselina (UA) različita kod sva četiri klastera. Indeks tjelesne težine (BMI) je kod klastera 2 i 3 vrlo sličan, dok klaster 1 i 4 odskaču.

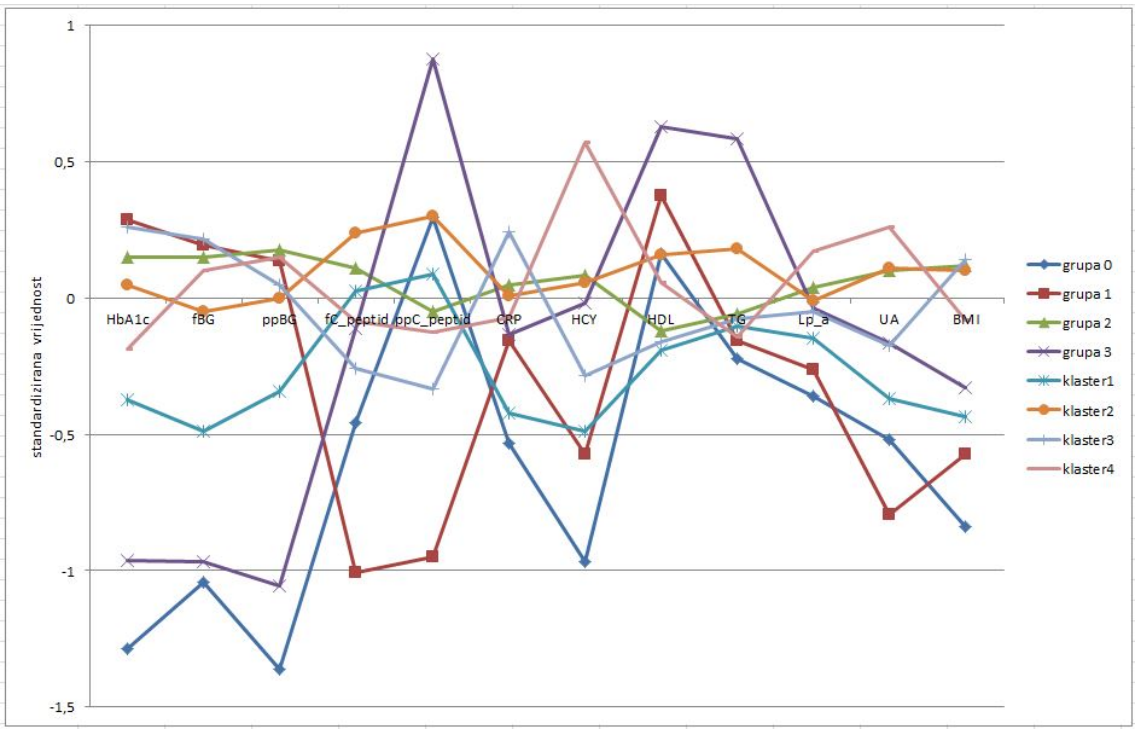
Da bih dobila tablicu koja prikazuje kako su pacijenti iz određene grupe raspoređeni po klasterima koristim naredbu:

```
proc freq data=Clust;
tables Groups*Cluster;
run;
```

Frequency Percent Row Pct Col Pct	Table of Groups by CLUSTER				
	CLUSTER(Cluster)				
Groups	**1**	**2**	**3**	**4**	**Total**
0	3 2.11 75.00 15.79	0 0.00 0.00 0.00	1 0.70 25.00 2.50	0 0.00 0.00 0.00	4 2.82
1	2 1.41 22.22 10.53	3 2.11 33.33 5.77	4 2.82 44.44 10.00	0 0.00 0.00 0.00	9 6.34
2	13 9.15 11.40 68.42	42 29.58 36.84 80.77	32 22.54 28.07 80.00	27 19.01 23.68 87.10	114 80.28
3	1 0.70 6.67 5.26	7 4.93 46.67 13.46	3 2.11 20.00 7.50	4 2.82 26.67 12.90	15 10.56
Total	19 13.38	52 36.62	40 28.17	31 21.83	142 100.00
Frequency Missing = 118					

Slika 2.12: Podjela grupa po klasterima (ispis iz SAS-a)

Iz tablice na slici 2.12 vidi se da ima malo potpunih podataka iz grupa 0, 1 i 3, te oni vjerojatno neće jako utjecat na dobivene klastere. Pacijenti iz grupe 2 su većinom stavljeni u klaster 2, međutim određeni postotak njih se nalazi i u ostalim dobivenim klasterima.



Slika 2.13: Aritmetičke sredine stvarnih grupa i 4 dobivena klastera po analiziranim varijablama

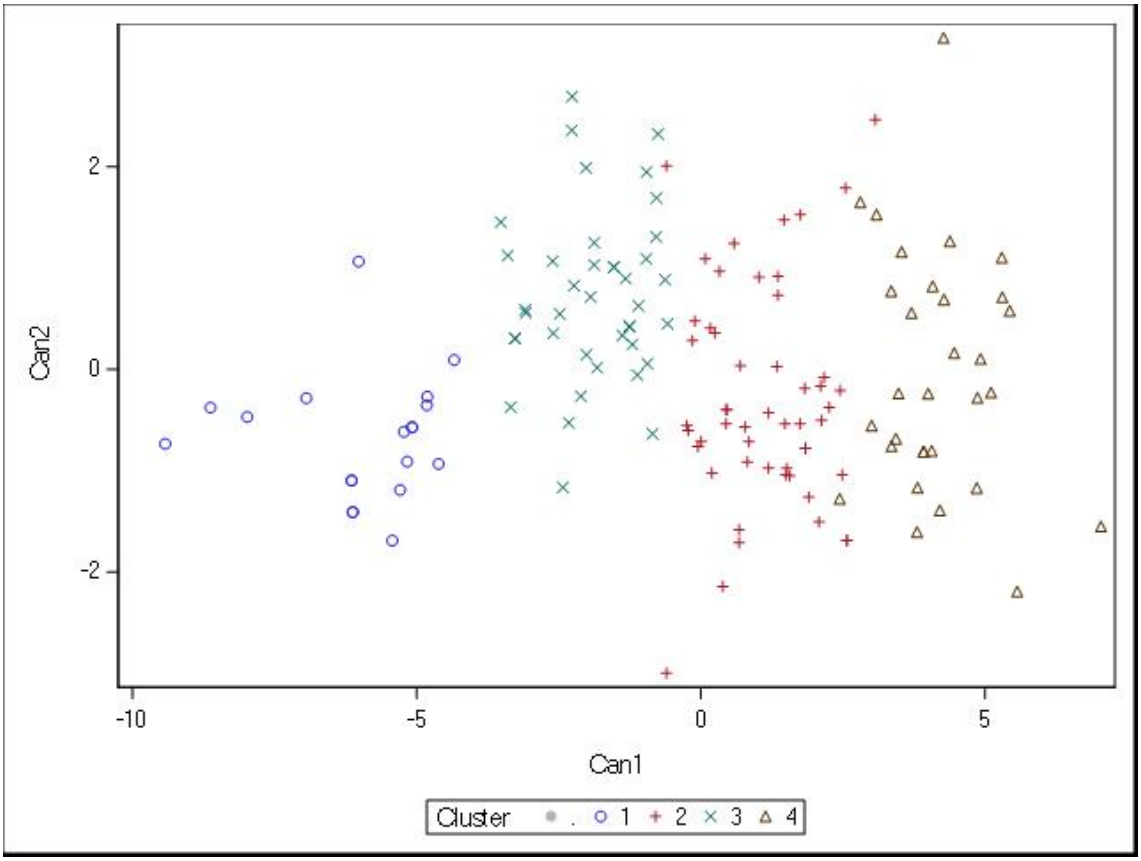
Iz prikaza aritmetičkih sredina stvarnih grupa te dobivena 4 klastera na slici 2.13 vidi se da niti jedna klaster ne prati pripadnu grupu u potpunosti što potvrđuje i točnosti klasifikacije na slici 2.12. Unatoč tome, uočava se da su aritmetičke sredine grupe 2 i klastera 2 najbliže, što je vidljivo iz tablice na slici 2.12 jer je najviše pacijenata iz grupe 2 raspoređeno u klaster 2, (36.84%). Također se vidi sličnost između aritmetičkih sredina varijabli grupe 1 i klastera 3, jer je 44.44% pacijenata iz grupe 1 stavljeno u klaster 3. Klaster 4 ne može se povezati niti s jednom grupom, jer je u njega raspoređen otprilike jednak postotak pacijenata iz grupa 2 i 3. U grupi 0 su samo 4 pacijenta s potpunim podacima, te oni ne utječu niti na jedan klaster.

Ako se u klasterskoj analizi koriste više od dvije varijable, tada se za grafički prikaz klastera koristi procedura `proc candisc`.

SAS kod:

```
proc candisc data=Clust out=Can noprint;
class Cluster;
```

```
var Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a UA BMI;
run;
proc sgplot data=Can;
scatter y=Can2 x=Can1 / group=Cluster;
run;
```



Slika 2.14: Grafički prikaz četiri klastera

Na slici 2.14 vidi se da je klaster 1 najviše odvojen od ostalih kod kojih se uočavaju preklapanja, te se ne mogu povući jasne granice među klasterima 2, 3 i 4.

S obzirom da pacijenata u grupi 3 ima najmanje, te je to zapravo međugrupa između zdravih pacijenata i onih s dijabetesom, a hijerarhijsko klasteriranje je u dva slučaja (Gowe-

rova mjera udaljenosti i Wardova metoda klasteriranja, te Manhattan mjera udaljenosti i metoda maksimuma) pacijente rasporedilo u tri klastera, provela sam i nehijerarhijski algoritam k-srednjih vrijednosti za k=3.

SAS kod:

```
proc fastclus data=standa out=Clust maxclusters=3 nomiss maxiter=300;
var Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a UA BMI;
run;
```

Cluster Means

Cluster	Dob	HbA1c	fBG	ppBG	fC_peptid	ppC_peptid	CRP	HCY
1	38.80000000	-0.16584859	-0.23907425	-0.14435804	-0.20785951	-0.20571016	-0.34124098	-0.43598739
2	56.98461538	0.17791449	0.06873312	-0.01332464	0.10947884	0.01978794	0.08809322	-0.16359684
3	71.51923077	-0.14265822	0.02902314	0.08605871	-0.03691609	0.07416419	0.05394164	0.41410538

Cluster Means

Cluster	HDL	TG	Lp_a	UA	BMI
1	-0.18038151	-0.12685760	-0.06763292	-0.46563153	-0.55593373
2	-0.09833517	-0.02254973	-0.03485656	0.04735409	0.31050999
3	0.20964084	0.08917640	0.07608652	0.16466870	-0.12086166

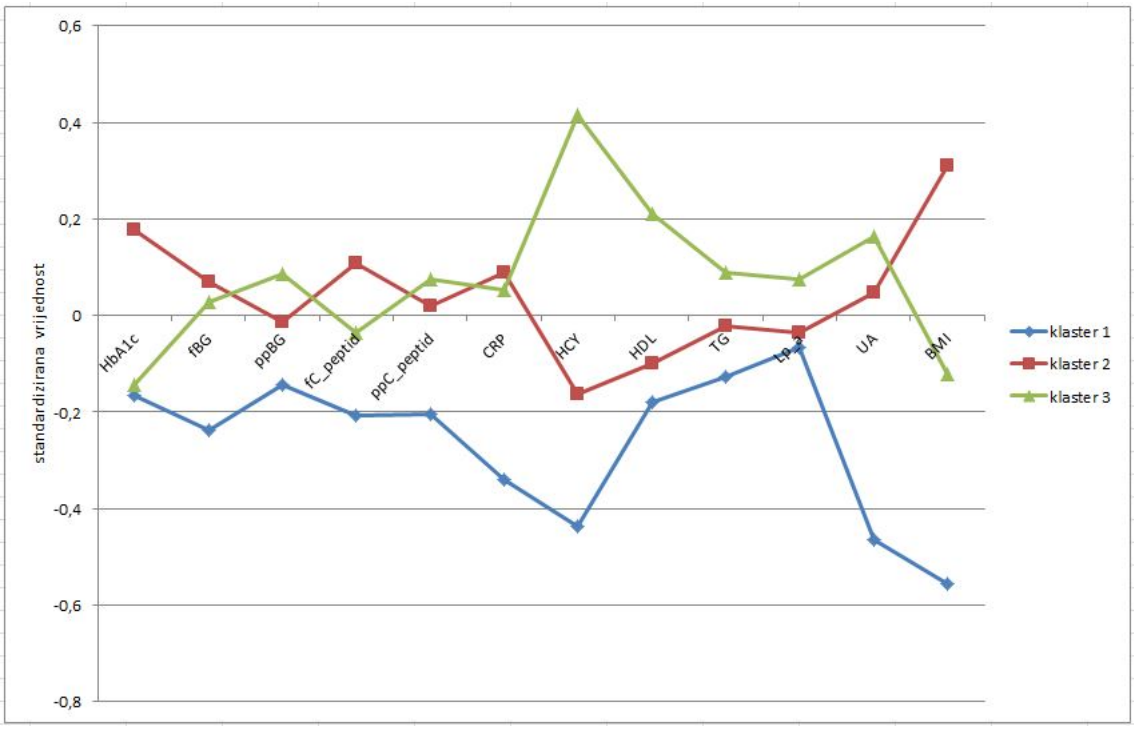
Cluster Standard Deviations

Cluster	Dob	HbA1c	fBG	ppBG	fC_peptid	ppC_peptid	CRP	HCY
1	6.952217872	1.071536922	0.915544773	1.197156319	0.932115113	1.224261937	0.344860285	0.633930355
2	4.790916887	0.967090761	1.000908606	0.882294918	1.129240257	0.948088735	0.940935898	0.498965776
3	4.932844642	0.987997376	1.038348255	1.046614771	0.847210605	0.951069323	1.236026075	1.400366809

Cluster Standard Deviations

Cluster	HDL	TG	Lp_a	UA	BMI
1	0.693011934	0.190631508	0.815086118	0.698125792	0.955568628
2	1.035144759	0.499862854	1.148478589	0.967766993	1.107696676
3	1.058751677	1.554800963	0.887214956	1.106009653	0.711317797

Slika 2.15: Aritmetičke sredine i standardne devijacije varijabli po klasterima (ispis iz SAS-a)



Slika 2.16: Aritmetičke sredine 3 dobivena klastera po analiziranim varijablama

Iz grafičkog prikaza aritmetičkih sredina 3 dobivena klastera na slici 2.16 prvo se uočava da se klaster 1 izdvaja, sve varijable mu imaju ispodposječne vrijednosti. Klasteri 2 i 3 imaju slične vrijednosti za glukozu natašte (fBG), ppC peptid te CRP, a razlika se javlja kod homocisteina (HCY), HDL-a, triglicerida (TG) i lipoproteina koje klaster 3 ima iznadprosječne, a klaster 2 ispodprosječne. Slika 2.16 također sugerira da najveću razliku među klasterima rade varijable homocitein, mokraćna kiselina (UA) i indeks tjelesne mase (BMI).

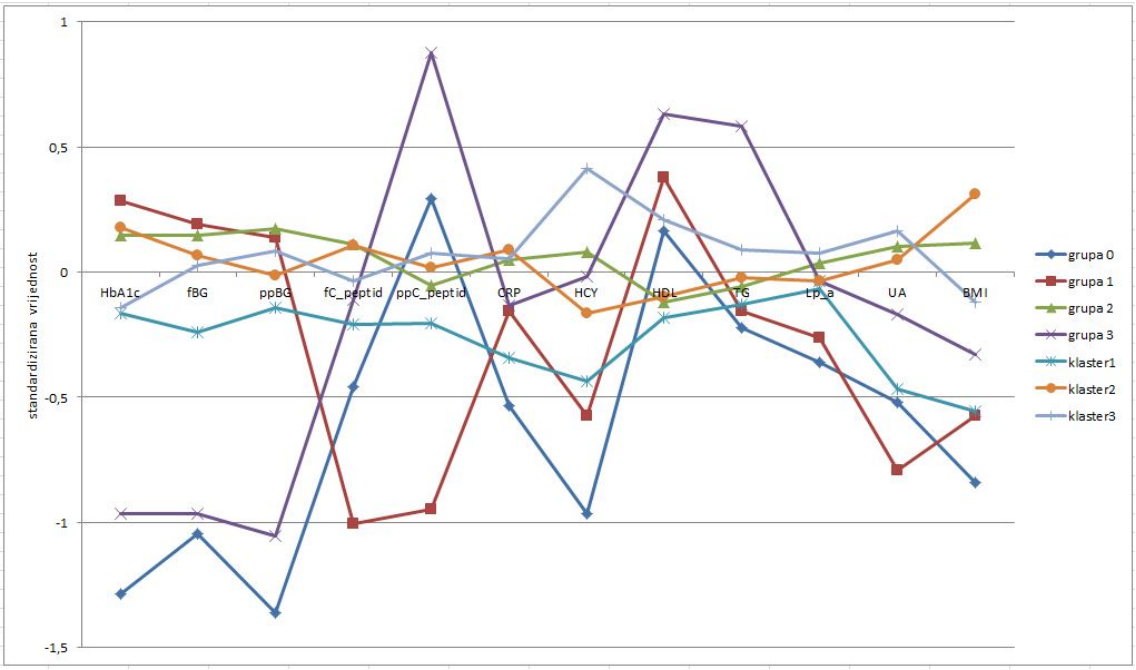
Za prikaz tablice frekvencije točnosti klasifikacije odnosno raspodjelu grupa po klasterima koristim SAS kod:

```
proc freq data=Clust;
tables Groups*Cluster;
run;
```

Frequency Percent Row Pct Col Pct	Table of Groups by CLUSTER			
	CLUSTER(Cluster)			
Groups	1	2	3	Total
0	3 2.11 75.00 12.00	1 0.70 25.00 1.54	0 0.00 0.00 0.00	4 2.82
1	3 2.11 33.33 12.00	5 3.52 55.56 7.69	1 0.70 11.11 1.92	9 6.34
2	17 11.97 14.91 68.00	53 37.32 46.49 81.54	44 30.99 38.60 84.62	114 80.28
3	2 1.41 13.33 8.00	6 4.23 40.00 9.23	7 4.93 46.67 13.46	15 10.56
Total	25 17.61	65 45.77	52 36.62	142 100.00
Frequency Missing = 118				

Slika 2.17: Podjela grupa po klasterima

Iz tablice na slici 2.17 vidi se da ima malo potpunih podataka iz grupa 0, 1 i 3, te oni vjerojatno neće utjecati na dobivene klastere. Pacijenti iz grupe 2 su većinom stavljeni u klaster 2, međutim određeni postotak njih se nalazi i u ostalim dobivenim klasterima.

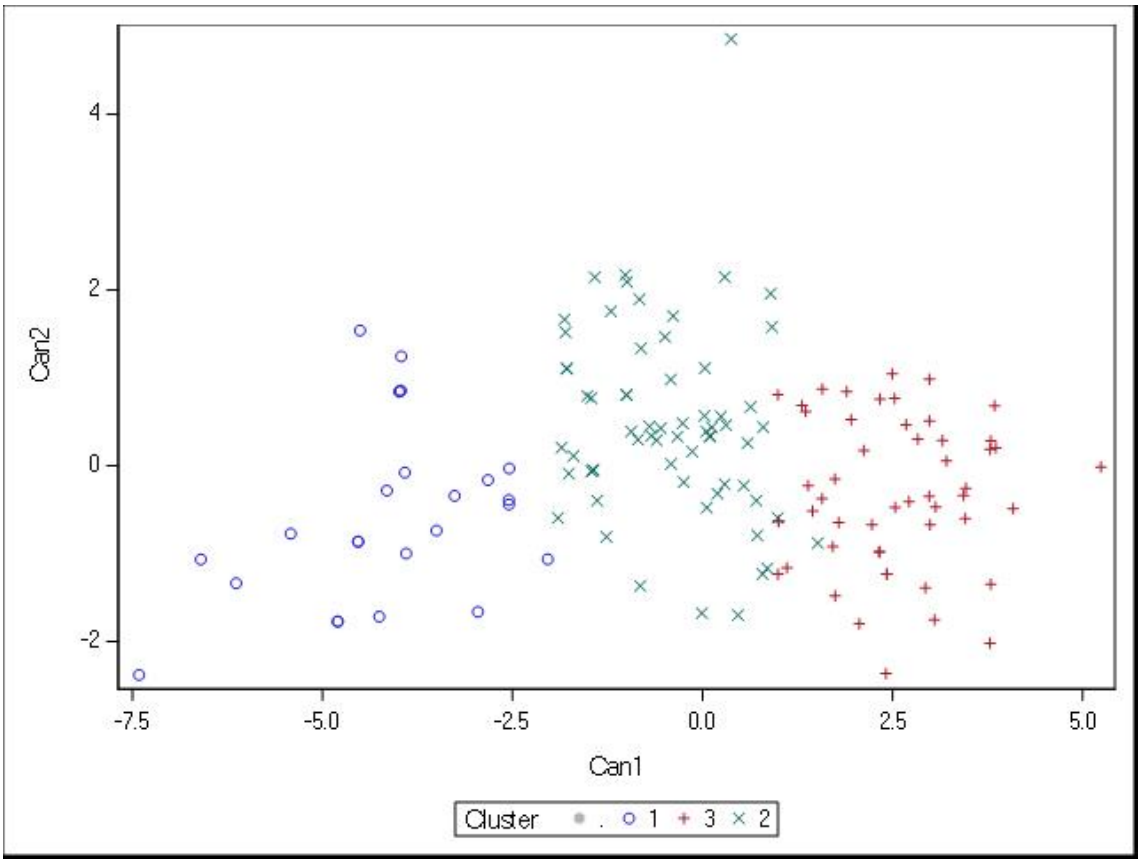


Slika 2.18: Aritmetičke sredine stvarnih grupa i 3 dobivena klastera po analiziranim varijablama

Iz prikaza aritmetičkih sredina stvarnih grupa te 3 dobivena klastera na slici 2.18 vidi se da niti jedna klaster ne prati neku grupu u potpunosti. Unatoč tome, uočava se da su aritmetičke sredine varijabli grupe 2 i klastera 2 najbliže što potvrđuje i tablica na slici 2.17, jer je najviše pacijenata iz grupe 2 raspoređeno u klaster 2 (46.49%). Izuzevši prve tri prikazane varijable, klaster 3 otprilike prati grupu 3, što ne iznenađuje s obzirom da je $46,64\%$ pacijenata iz grupe 3 stavljeno u klaster 3. Kod klastera 1 su aritmetičke sredine svih varijabli ispodprosječne, a u njega su raspoređeni pacijenti iz sve četiri grupe. U grupi 0 su samo 4 pacijenta s potpunim podacima, te oni ne utječu jako niti na jedan klaster.

Za prikaz dobivenih klastera u koordinatnom sustavu korisi se sljedeći SAS kod:

```
proc candisc data=Clust out=Can noprint;
class Cluster;
var Dob HbA1c fBG ppBG fC_peptid ppC_peptid CRP HCY HDL TG Lp_a UA BMI;
run;
proc sgplot data=Can;
scatter y=Can2 x=Can1 / group=Cluster;
run;
```



Slika 2.19: Grafički prikaz tri klastera

Na slici 2.19 vidi se da se elementi iz klastera 2 i 3 na jednom djelu preklapaju dok su oni iz klastera 1 više odvojeni. To je sugerirala i slika 2.16 gdje se vidi da aritmetičke sredine svih varijabli iz klastera 1 imaju ispodprosječne vrijednosti, dok većina varijabli iz klastera 2 i 3 ima prosječne ili iznadprosječne.

Za daljnju razradu i tumačenje bila bi neophodna pomoć i stručnost liječnika.

Bibliografija

[1] *SAS/STAT® 9.2 User's Guide The CLUSTER Procedure*. dostupno na https://support.sas.com/documentation/cdl/en/statugcluster/61777/PDF/default/statugcluster.pdf (travanj 2016.).

[2] *SAS/STAT® 9.2 User's Guide The DISTANCE Procedure*. dostupno na https://support.sas.com/documentation/cdl/en/statugdistance/61780/PDF/default/statugdistance.pdf (travanj 2016.).

[3] Ferligoj, A.: *Cluster analysis*, 2002. materijali sa School of Biometrics, Cavtat, June 25, 2002.

[4] Hastie, T., R. Tibshirani i J.H. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[5] Lin, H.: *Clustering*. dostupno na http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf (travanj 2016.).

[6] Periklis, A.: *Data Clustering Techniques*. Rapport technique, University of Toronto, Department of Computer Science, 2002.

Sažetak

U ovom radu opisana je klasterska analiza te njena primjena na bazu dijabetičara podijeljenih prema stupnju bolesti. Klasterska analiza podrazumijeva grupiranje jedinki u klastere tako da su jedinke unutar klastera ”najsličnije”. Kod klasteriranja prvo moramo definirati udaljenost tj. sličnost između dvije jedinke, a nakon toga definiramo algoritam grupiranja. Postoji hijerarhijsko i nehijerarhijsko klasteriranje.
U opisanom primjeru mjereno je 13 varijabli na osnovu kojih je bilo očekivano da će se pacijenti grupirati u četiri klastera (zdravi, dijabetes tipa I, dijabetes tipa II, predijabetes). Napravljeni su hijerarhijski klasterinzi i to pomoću euklidske mjere udaljenosti i metode prosjeka, Gower mjere udaljenosti i Wardove metode klasteriranja, te Manhattan mjere udaljenosti i metode maksimuma, kao i nehijerarhijski klasterinzi algoritmom k-srednjih vrijednosti za k=4 i 3.

Summary

This paper describes cluster analysis and its application to the base of diabetics divided according to degree of the disease. Cluster analysis involves grouping elements into clusters so the elements within one cluster are the most similar. For clustering we must first define the distance (similarity) between two elements and then define the grouping algorithm. There are hierarchical and nonhierarchical clustering.
In described example there were 13 variables measured from which was expected that patients will be grouped into four clusters (healthy, type I diabetes, type II diabetes, prediabetes). Hierarchical clustering was made with the euclidian distance and the average method, the Gower distance and the Ward method and the Manhattan distance and the maximum method. Also, nonhierarchical clustering was made with k-means algorithm for k=4 and 3.

Životopis

Rođena sam 20. lipnja 1992. godine u Zagrebu. Završila sam OŠ Vrbani te V. gimnaziju. 2011. godine upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu kojeg završavam 2014. te upisujem diplomski sveučilišni studij Matematička statistika na istom fakultetu.